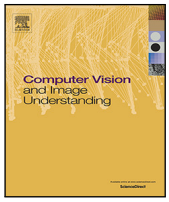
Computer Vision and Image Understanding 237 (2023) 103833

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



On learning distribution alignment for video-based visible-infrared person re-identification

Pengfei Fang [a,b,*], Yaojun Hu [c], Shipeng Zhu [a,b], Hui Xue [a,b]

[a] School of Computer Science and Engineering, Southeast University, Nanjing, 210096, China
[b] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[c] College of Science, Zhejiang University of Technology, Hangzhou 310023, China

ARTICLE INFO

Communicated by Juergen Gall

MSC:
41A05
41A10
65D05
65D17

Keywords:
Visible-infrared person re-identification
Video
Discrete distribution
Optimal transport
Symmetric optimal transport

ABSTRACT

This paper studies the matching problem of cross-modality video data from a discrete distribution alignment view. Central to this discussion is the visible-infrared person re-identification (VI-reID), a crucial feature that bolsters surveillance systems' efficacy in monitoring individuals across diverse lighting conditions. Going beyond traditional image-to-image matching paradigms, a recent study shows that temporal information can bring richer cues to encode the pedestrian representation, improving the representation power of deep neural networks. However, this integration further complicates cross-modality data matching due to the joint processing of spatial and temporal information. This paper formulates the video data as a discrete distribution and aligns the cross-modality video representation by reducing the matching cost between the two distributions. To this end, a natural idea for aligning the videos is to reduce the divergence of distributions. Moreover, the powerful optimal transport (OT) scheme, which generates the optimal matching flows and establishes the relevance of two distributions, is also employed as a way to measure the distance of distributions. Nevertheless, we observe that endowing the OT in the advanced VI-reID feature extractor leads to a non-symmetric measurement. To mitigate this, the paper introduces a new metric, namely symmetric optimal transport (SOT), reformulating OT into a symmetric form. Thorough analyses and empirical studies affirm the superiority of the proposed SOT, which significantly outperforms the current state-of-the-art methods according to standard benchmarking evaluations.

1. Introduction

This paper studies the person re-identification (reID) problem, particularly interested in the matching problem of the cross-modality video data from the distribution alignment view.

Person reID is an essential problem in video surveillance systems, and it aims to search for target pedestrians in a large database (Gong et al., 2014; Cho et al., 2019; Li et al., 2021; Liu et al., 2023b). This requires deep neural networks (DNNs) to create useful person embeddings, such that the embedding space can reveal the underlying distribution of pedestrian data in raw image spaces, i.e., small intra-class variance and large inter-class variance (Ye et al., 2021b; Zheng et al., 2016). In the past few years, the community has made significant progress for the person re-ID with homogeneous data (Ye et al., 2021b; Zheng et al., 2016). These algorithms train the re-ID machine in the metric learning paradigm (Weinberger and Saul, 2009; Schroff et al., 2015; Wang et al., 2017) and produce discriminative person representations in a homogeneous embedding space. However, it still remains a gap in successfully deploying those methods in complicated real-world scenarios, where it requires the machine to make robust inference for matching the pedestrian images captured by RGB and infrared cameras.

By formulating the task as a cross-modality data matching problem, the visible-infrared person reID (VI-reID) is proposed to learn a joint embedding space for the visible images and infrared images, making it possible to surveillance in 24 h (Wu et al., 2017). It is not an easy task given the challenge of large modality discrepancy between visible images and infrared images (Wu et al., 2017), as well as the misalignment issue-inherited in the re-ID task (Fang et al., 2019; Suh et al., 2018). Specifically, the modality discrepancy is caused by the diverse imaging devices and the misalignment in this task refers to the situation where feature maps are misaligned due to spatial nuances, e.g., movements of body parts, pose, background, etc. Much effort has gone into addressing these issues. In particular, two schools of image-to-image retrieval methods are mainly studied, including representation-based methods (Ye et al., 2020; Tian et al., 2021; Chen et al., 2021; Hao et al., 2021; Fu et al., 2021; Wei et al., 2020; Gong et al., 2023; Liu et al., 2023a) and generative-based methods (Wang et al., 2019a,b, 2020a; Choi et al., 2020; Ye et al., 2021a; Wei et al.,

* Corresponding author at: School of Computer Science and Engineering, Southeast University, Nanjing, 210096, China.
E-mail addresses: fangpengfei@seu.edu.cn (P. Fang), yaojunhu@zjut.edu.cn (Y. Hu), shipengzhu@seu.edu.cn (S. Zhu), hxue@seu.edu.cn (H. Xue).

https://doi.org/10.1016/j.cviu.2023.103833
Received 27 June 2023; Received in revised form 7 August 2023; Accepted 9 September 2023
Available online 21 September 2023

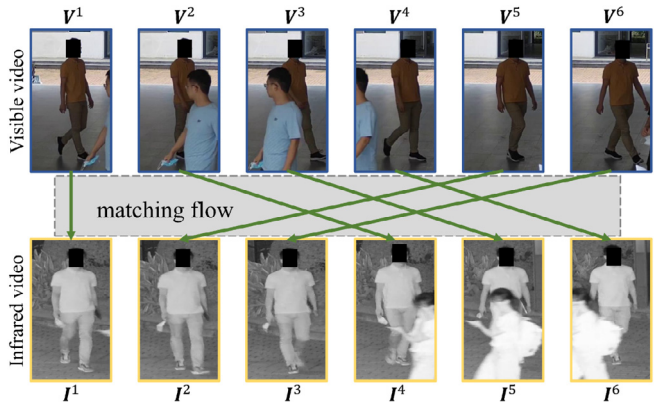


Fig. 1. The misalignment issue in the video matching task. Given two video sequences, i.e., one visible video and one infrared video, the factors of pose, occlusion, etc. cause the misalignment issue of matching videos.

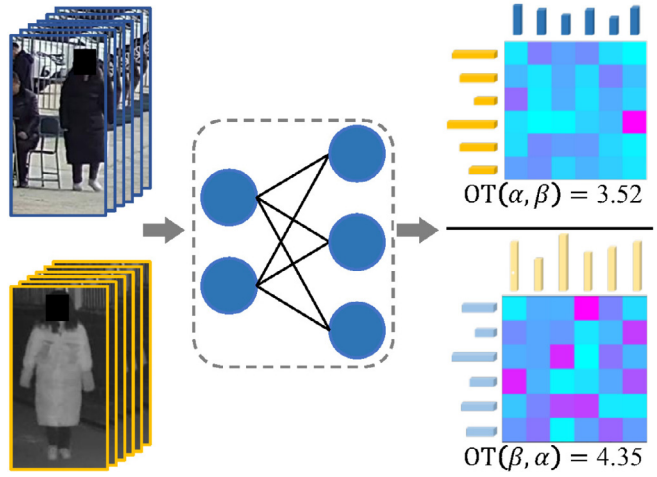


Fig. 2. Given two video sequences, the corresponding frame features can be encoded by the deep neural network, and such frame features can be modeled as a discrete distribution. Then one can calculate the $\mathrm{OT}(\alpha,\beta)$ and $\mathrm{OT}(\beta,\alpha)$. Endowing the OT in the advanced VI re-ID feature extractor leads to a non-symmetric measurement.

2021a). The representation-based methods develop advanced metric learning approaches or neural architectures to facilitate the alignment of two embedding spaces, while generative-based methods leverage generators to transfer the pedestrian images from one modality to another, such that the retrieval procedure can be realized in a common embedding space.

Despite the promising achievement made in the existing works, it is still far more to address the VI-reID task using the image-to-image retrieval paradigm, due to infrared images cannot provide informative appearance features, e.g., color, texture, etc. For example, different people with similar wearing can be easily recognized as the same person in infrared images. One possible solution is to leverage the video data, i.e., the image sequences, which offers temporal information on top of the spatial appearance information. This enables the neural network to fuse the motion feature, jointly learned from the temporal and spatial information, into the person embeddings, improving its discrimination. This is first studied by Lin et al. in Lin et al. (2022). This work establishes the first video-based visible-infrared pedestrian dataset, named HITSZ-VCM, as a benchmark for the video-based VI-reID task. On top of the benchmark, Lin et al. further propose a modal-invariant and temporal-memory learning (MITML) scheme to reduce the modality variance, resulting in improvement over the baselines.

The work (Lin et al., 2022) employs the popular pipeline to produce the video-level embedding, first extracting the frame-level features and then aggregating them to a video-level feature. That said, even though the video-level feature contains rich temporal information, the aggregation operation will result in information loss of the original frame features. To make use of the information flow in the video data, a potential solution is to learn a matrix representation per video, gaining enough information to represent the video data. Defining a metric to measure the distance of two matrix representations is not without difficulties and this paper addresses this issue by matching two distributions. In doing so, this paper models the video data, i.e., matrix representation, as a discrete distribution and realizes the video matching problem in a discrete distribution alignment view. A natural idea to align distributions is to minimize the divergence, e.g., Kullback–Leibler divergence, Jensen–Shannon divergence, etc, between distributions. However, such a "hard" measurement scheme cannot determine the optimal matching flow between two distributions, which may lead to misalignment of the video matching. Also, the formulation of such divergences inherently causes numerical instability.

The optimal transport (OT) has gained interest in the learning community as it constructs an alternative concept of distance measurement between distributions and it benefits a diverse set of learning scenarios, including visual understanding (Zhang et al., 2020), image generation (Arjovsky et al., 2017), biological data analysis (del Barrio et al., 2020), text understanding (Zhao et al., 2021), etc. OT defines the distance by formulating the transportation problem, and the optimal transportation plan can be determined by optimizing the linear programming problem. In this context, the determined transportation plan yields an optimal matching flow of two distributions, which results in the minimum matching cost. This motivates to match the video data via optimal transport scheme, such that the misalignment issue between videos can be mitigated (see Fig. 1). Specifically, minimizing the video matching cost can identify both the optimal matching flow and the parameters for neural networks. However, we observe that optimization via an OT loss cannot yield the optimal matching flow between two distributions. This is shown in Fig. 2. Given two video sequences, the corresponding frame features can be encoded by the deep networks. Such frame features can be modeled as a discrete distribution. This allows to calculate the transportation plan between two distributions, e.g., a and b. However, the transportation plan for $\mathrm{OT}(\alpha,\beta)$ and $\mathrm{OT}(\beta,\alpha)$ is non-symmetric[1] (see Fig. 2), leading to unstable performance in the inference stage. This paper formulates a symmetric optimal transport (SOT) scheme to address this issue. The **contributions** of this paper are as follows:

- This paper formally formulates the video sequences into a discrete distribution and empirically evaluates the effectiveness of well-established distribution measurements to align the video data.
- This paper further provides a good practice of distribution alignment by proposing a symmetric optimal transport (SOT) scheme.
- Thorough experiments are conducted to evaluate the effectiveness of the proposed SOT, which attains state-of-the-art performance on the public video VI-reID benchmark.

2. Related work

In this section, we briefly discuss the related work, including person re-identification and optimal transport.

[1] The optimal transport can be understood as a metric to calculate the optimal matching cost of two distributions. In other words, the optimal matching cost for two video sequences should be symmetric, e.g., $\mathrm{OT}(\alpha,\beta) = \mathrm{OT}(\beta,\alpha)$.

2.1. Person re-identification

As an important component in the intelligent surveillance system, person reID has made significant improvement in the past few years. The main goal of person reID is to create a generalizable embedding space, such that the retrieval task can be performed well for unseen persons in the inference stage (Ye et al., 2021b; Zheng et al., 2016). In the deep learning ear, the community leverages the convolutional neural network (CNN) or vision transformer (ViT) to extract the person's appearance features and develops algorithms to improve the representation power of the backbones. Some algorithms focus on mining the relationship of person images and develop loss functions to constrain the optimization process (Hermans et al., 2017; Zhou et al., 2017). Considering the content of person images, some developments employ auxiliary information, including human poses, human attributes, and visual attention, as cues to boost the representation power of images (Fang et al., 2019; Tay et al., 2019; Su et al., 2017; Fang et al., 2021).

Apart from the setting of intra-modality reID, the surveillance system is required to address the reID task in complex scenarios (Wang et al., 2020b). Such complex scenarios contain different camera specifications (e.g., low-resolution vs. high-resolution image), different sensory devices (e.g., infrared light devices vs. visible light devices), and different data formats (e.g., text description vs. digital images). This paper particularly focuses on the setting of visible-infrared image reID. This setting aims to learn a modality-invariant embedding space, such that the same identity, described by heterogeneous images, can be clustered. To this end, two learning paradigms are adopted to learn a common embedding space for heterogeneous images.

One learning paradigm explicitly learns the modality-specific or modality-shared features of the person images. In Wu et al. (2017), Wu et al. augment the network to learn domain-specific features via zero-padding to visible or infrared images. Constrained by a cross-modality similarity preservation loss, the work in Wu et al. (2020) enables the network to learn domain-shared features. The following work (Lu et al., 2020) proposes the cross-modality shared-specific feature transfer algorithm (cm-SSFT) to benefit from both modality-specific and modality-shared features jointly. Works in Li et al. (2020) and Wei et al. (2021b) propose to learn a middle modality, bridging the modality gap of the images. The part features, developed for the traditional reID task (Sun et al., 2018), is also effectiveness to present different modalities of person images (Wei et al., 2021a). Another learning paradigm transfers the modality of the input image to another, such that the matching can be realized in the same modality space. The AlignGAN translates visible images to infrared images and trains the network to align the in both pixel level and feature level (Wang et al., 2019a). Wang et al. adopt a similar idea to generate two modalities of images in the same latent space (Wang et al., 2019b). The following attempts either to produce the light conditions to rich person appearance features (Wang et al., 2020a) or disentangles the light conditions and only encode person-discriminative factors (Choi et al., 2020).

Compared to the image level VI-reID, matching the pedestrian using video data makes this task more difficult (Lin et al., 2022). The first attempt develops temporal memory refinement (TMR) to capture the motion information of pedestrian video, and fuses it in the pedestrian embedding (Lin et al., 2022). In contrast to Lin et al. (2022), we address the video VI-reID task in the distribution alignment view.

2.2. Optimal transport

Optimal transport (OT), first proposed by Gaspard Monge in 1781 (Monge, 1781), is proposed to study the allocation of resources. It is then formulated as a general problem of efficiently moving one distribution mass to another. This, later, motivates the learning community to explore the benefits of OT as a distance metric in various artificial intelligence (AI) applications, e.g., computer vision, and natural language processing, to name but a few.

In the context of computer vision, OT can yield the optimal matching flow of two images, avoiding the misalignment issue. An early attempt first represents the image as a color histogram, which can be modeled as a distribution. The OT can then be employed as the distance metric to retrieve images (Rubner et al., 2000). Instead of using the color histogram as image representation, the semantic features, extracted by DNNs, can also represent images in the deep learning ear. In doing so, the works (Zhang et al., 2020; Liu et al., 2020) model the feature map as a distribution, and leverages to address the few-shot learning and dense correspondence problems. In Solomon et al. (2014), an improved OT is used to calculate the distance of the discrete surfaces. The OT can also improve the generative models, due to OT can optimize the generator to produce data distribution like that of the training data. This is first studied in Arjovsky et al. (2017), where the generative adversarial network (GAN) adopts the Wasserstein distance, benefiting the stability of learning and avoiding the collapse issue. This idea is further elaborated in Salimans et al. (2018), Genevay et al. (2017) and Bellemare et al. (2017), and improves the learning behavior of GAN models. Considering the document representation as a distribution over words, Yurochkin et al. propose hierarchical optimal transport to measure the distance between documents (Yurochkin et al., 2019). On top of the document representation, Zhao et al. utilize OT to optimize neural topic models (NTM), leading to improved document and topic representations jointly (Zhao et al., 2021). In Xu et al. (2019), a new learning framework, namely, Gromov–Wasserstein Learning, adopts the OT as a regularizer to establish correspondence between graphs. In contrast to Xu et al. (2019), Kolouri calculates the OT between a reference distribution to each node embedding, leading to an improved and fast graph embedding learning network (Kolouri et al., 2021). Having its efficiency in mind, this paper formally models the video sequence as a discrete distribution, such that one can yield the video matching issue via OT.

3. Method

This section details the proposed method: starting from the problem formulation of the video-based VI-reID task, followed by the network architecture as the video feature extractor. We then formally formulate the video sequences as a discrete distribution. Thereafter, we discuss the well-established distribution measurements and present the proposed method.

Notations. Formally, we use $\mathbb{R}^n$, $\mathbb{R}^{m\times n}$, $\mathbb{R}^{c\times m\times n}$ and $\mathbb{R}^{t\times c\times m\times n}$ to denote n-dimensional Euclidean spaces, the space of $m\times n$ real matrices, the image and video spaces. The Dirac delta function δ is defined as

$$\delta_{\boldsymbol{x}} = \begin{cases} 0, & \boldsymbol{x}\neq 0 \\ \infty, & \boldsymbol{x}=0 \end{cases}, \tag{1}$$

where $\int_{-\infty}^{\infty}\delta_{\boldsymbol{x}}d\boldsymbol{x}=1$. For a set of parameters, m-simplex represents the simplest possible polytope in m dimension, defined as $\Delta_{m-1}=\{\boldsymbol{s}\in\mathbb{R}_+^m : \sum_{i=1}^m s^i=1\}$. $\mathbb{1}_m\in\mathbb{R}^m$ indicates an all-ones vector, where all elements are 1.

3.1. Problem formulation

Let two fourth-order tensor, $\mathcal{V}_r=[\boldsymbol{V}_r^1,\boldsymbol{V}_r^2,\ldots,\boldsymbol{V}_r^T]\in\mathbb{R}^{T\times C\times H\times W}$ and $\mathcal{I}_r=[\boldsymbol{I}_r^1,\boldsymbol{I}_r^2,\ldots,\boldsymbol{I}_r^T]\in\mathbb{R}^{T\times C\times H\times W}$, denote the rth visible video sequence and infrared video sequence of a pedestrian, where T, C, H and W are the number of frames, channels, height and width of a video sequence, respectively. The training set contains M visible video sequences and N infrared video sequences, described by $\mathbb{V}=\{\mathcal{V}_r,y_r\}_{r=1}^M$ and $\mathbb{I}=\{\mathcal{I}_r,y_r\}_{r=1}^N$. The video-based VI-reID model, $f(\cdot|\theta)$, aims to learn a feature extractor, that projects two modalities of video data to the same embedding space, such that one can yield the task of cross-modality video retrieval. In doing so, a proper loss function, $\mathcal{L}$, should be adopted as an objective to optimize the parameters of the feature extractor. In this paper, we model the video sequence in the embedding space as a discrete distribution and hence, we are interested in aligning discrete distributions, as a contribution to the video VI-reID community.

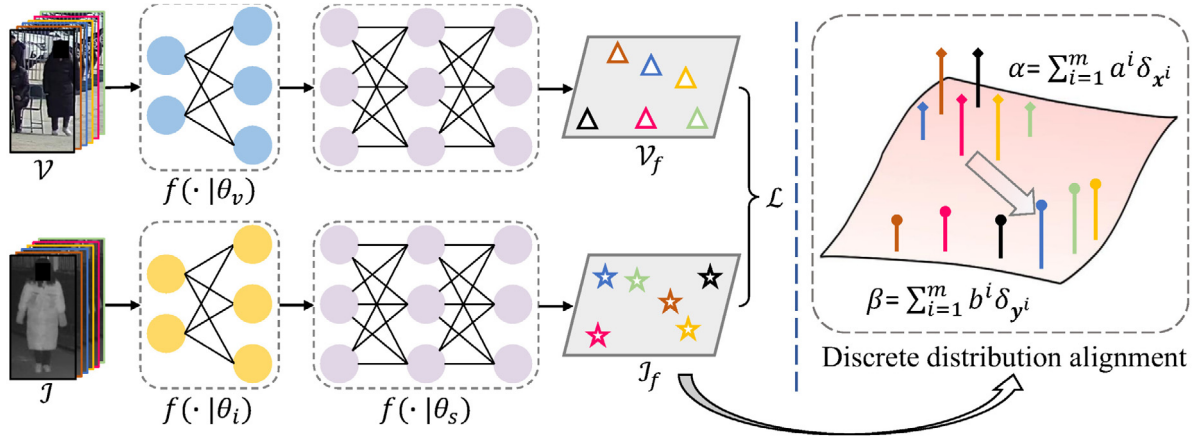


Fig. 3. The pipeline of cross-modality video matching in discrete distribution alignment view. Given visible video $\mathcal{V}$ and infrared video $\mathcal{I}$ as input, the modality learning network, $f(\cdot|\theta_v)$ and $f(\cdot|\theta_i)$, encode the modality features for the visible video and infrared video. Then a semantic learning network, $f(\cdot|\theta_s)$, is used to encode the semantic features for both modalities of data, resulting in representations of two videos, $\mathcal{V}_f$ and $\mathcal{I}_f$. We model the $\mathcal{V}_f$ and $\mathcal{I}_f$ as discrete distributions and realize the matching problem via distribution alignment.

3.2. Overview

We begin by providing a sketch of the DNN pipeline for the video VI-reID task, shown in Fig. 3. In the video VI-reID task, one ideal solution is to make use of a neural network to encode two modalities of videos into a joint embedding space, enabling the network to understand the modality-invariant video representations. In the training phase, two modalities of the video sequence are sampled as input to the neural network. The neural network contains two main components, e.g., a low-level modality learning network and a high-level semantic learning network. Specifically, given two video sequences, $\mathcal{V} = [\boldsymbol{V}^1, \boldsymbol{V}^2, \dots, \boldsymbol{V}^T]$ and $\mathcal{I} = [\boldsymbol{I}^1, \boldsymbol{I}^2, \dots, \boldsymbol{I}^T]$ with $\boldsymbol{V}^t, \boldsymbol{I}^t \in \mathbb{R}^{C\times H\times W}$, two low-level modality learning networks, $f(\cdot|\theta_v)$ and $f(\cdot|\theta_i)$, encode the modality features for the visible video and infrared video. Of note, $f(\cdot|\theta_v)$ and $f(\cdot|\theta_i)$ do not share the weights, such that the low-level modality learning networks can learn modality-specific features. A followed high-level semantic learning network, $f(\cdot|\theta_s)$, is further used to encode the modality-shareable feature for both modalities. This process can be formulated as $\mathcal{V}_f = f\big(f(\mathcal{V}|\theta_v)|\theta_s\big)$ and $\mathcal{I}_f = f\big(f(\mathcal{I}|\theta_i)|\theta_s\big)$, and produce a sequences of frame features, as $\mathcal{V}_f = [\boldsymbol{v}^1, \boldsymbol{v}^2, \dots, \boldsymbol{v}^T]$ and $\mathcal{I}_f = [\boldsymbol{i}^1, \boldsymbol{i}^2, \dots, \boldsymbol{i}^T]$, where $\boldsymbol{v}^t, \boldsymbol{i}^t \in \mathbb{R}^d$. We note that the feature of a video sequence can be fused to a compact video representation, and employ the loss functions the optimize the network. In our work, we want to leverage the matching cost of frame feature distributions as cues to optimize the network. The following section will formulate the frame features as a discrete distribution and discuss possible matching costs of distributions. Alone with the existing matching method, we also develop a new metric, termed symmetric optimal transport, as an alternative matching cost.

3.3. Distribution alignment

A common practice of learning video-level embedding consists of first extracting the frame-level features and then aggregating them to a video-level feature. The aggregation process inevitably causes information loss of the original frame features. Our work represents video data using a matrix to maximize information usage. This raises an issue of how to learn an embedding space for matrix representations and the misalignment for video matching. Thus we formally model the video representation as a discrete distribution and develop the distribution measurement to realize the metric learning paradigm for the matrix representations.

We first formally model the video sequence as a discrete distribution. Given a sequences of frame features, $\mathcal{X} = [\boldsymbol{x}^i]_{i=1}^m$ with $\boldsymbol{x}^i \in \mathbb{R}^d$, it can be assigned an empirical measurement on $\mathbb{R}^d$ by a discrete distribution, described as:

$$\alpha = \sum_{i=1}^{m} a^i \delta_{\boldsymbol{x}^i}, \tag{2}$$

where δ is the Dirac delta function. a^i is the weight for $\delta_{\boldsymbol{x}^i}$, satisfying that $\boldsymbol{a} = [a^i]_{i=1}^m \in \Delta_{m-1}$. This modeling allows using a discrete distribution to represent a video sequence, such that the video matching problem can be transferred to the distribution measurement problem.

In the following, we discuss some well-established measurements between distributions, e.g., $\alpha = \sum_{i=1}^m a^i \delta_{\boldsymbol{x}^i}$ and $\beta = \sum_{i=1}^m b^i \delta_{\boldsymbol{z}^i}$ for $\mathcal{X} = [\boldsymbol{x}^i]_{i=1}^m$ and $\mathcal{Z} = [\boldsymbol{z}^i]_{i=1}^m$ with $\boldsymbol{x}^i, \boldsymbol{z}^i \in \mathbb{R}^d$, and propose the symmetric optimal transport as an alternative to the distribution metric.

3.3.1. Kullback–Leibler divergence

The Kullback–Leibler divergence (KLD) is a type of statistical distance, which can quantify the match of two distributions. Its effectiveness has been used extensively as a measurement between distributions in the learning community (Zhang et al., 2018). In this paper, we also employ the KLD as a candidate to measure the matching cost of two distributions, e.g., α and β, which can be written as:

$$\mathrm{KLD}(\alpha \parallel \beta) = \sum_{i=1}^{m} a^i \log\left(\frac{a^i}{b^i}\right). \tag{3}$$

3.3.2. Jensen–Shannon divergence

As suggested in Eq. (3), the KLD is asymmetric. That said, KLD is not a valid metric for distributions. The Jensen–Shannon divergence (JSD) is further developed on top of the KLD and its symmetric property makes it a valid metric. It is also known as a total divergence to the average. The JSD is given by:

$$\begin{aligned}\mathrm{JSD}(\alpha \parallel \beta) &= \frac{1}{2}\big(\mathrm{KLD}(\alpha \parallel \beta) + \mathrm{KLD}(\beta \parallel \alpha)\big) \\ &= \frac{1}{2}\Big(\sum_{i=1}^{m} a^i \log\left(\frac{a^i}{b^i}\right) + \sum_{i=1}^{m} b^i \log\left(\frac{b^i}{a^i}\right)\Big).\end{aligned} \tag{4}$$

The JSD is symmetric since it holds that $\mathrm{JSD}(\alpha \parallel \beta) = \mathrm{JSD}(\beta \parallel \alpha)$. In practice, both the KLD and JSD suffer from the issue of numerical unstable when the denominator in Eqs. (3) and (4) is a tiny value.

3.3.3. Maximum mean discrepancy

The maximum mean discrepancy (MMD) is a kernel-based statistical measurement to determine the difference between two distributions.

As compared to KLD and JSD, MMD can enjoy the rich representation power of the kernel method. MMD is formulated as:

$$\begin{aligned}\mathrm{MMD}(\alpha,\beta) = &\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k(\boldsymbol{x}_i,\boldsymbol{x}_j)\\ &-\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k(\boldsymbol{x}_i,\boldsymbol{z}_j)+\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k(\boldsymbol{z}_i,\boldsymbol{z}_j),\end{aligned} \tag{5}$$

where $k(\cdot,\cdot)$ is a kernel function. In this paper, we use the popular RBF kernel, as $k(\boldsymbol{x},\boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{\sigma}\right)$.

3.3.4. Optimal transport

The optimal transport (OT) scheme is first formulated to identify the optimal transportation plan for resource allocation (Monge, 1781). It is then formulated to measure the distance between pairs of probability distributions (Arjovsky et al., 2017). In this part, we introduce Wasserstein (WS) distance for OT and the Sinkhorn algorithm to estimate the WS distance. Given two discrete distributions, e.g. $\alpha =$ and β, one can calculate WS distance as the optimal matching cost. Specifically, let $dist : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ be the distance metric defined on $\mathbb{R}^d$, the WS distance is given by:

$$\mathrm{OT}(\alpha,\beta) := W(\alpha,\beta) = \min_{\boldsymbol{P}\in\Gamma(\alpha,\beta)}\langle \boldsymbol{P},\boldsymbol{D}\rangle, \tag{6}$$

where $\boldsymbol{P}$ and $\Gamma(\alpha,\beta) = \{\boldsymbol{P}\in\mathbb{R}_+^{m\times m} : \boldsymbol{P}\mathbb{1}_m = \boldsymbol{a}, \boldsymbol{P}^\top\mathbb{1}_m = \boldsymbol{b}\}$ denote transportation plan and transportation polytope, respectively. In Eq. (6), $\boldsymbol{D}$ is the distance matrix with each element being $\boldsymbol{D} = [dist(\boldsymbol{x}_i,\boldsymbol{z}_j)]_{i=1,j=1}^{m,m}$. In this paper, the distance metric is specified as $dist(\boldsymbol{x}_i,\boldsymbol{z}_j) = 1-\cos(\boldsymbol{x}_i,\boldsymbol{z}_j)$, where $\cos(\cdot,\cdot)\in[-1,1]$ is the cosine similarity. Solving the optimization problem in Eq. (6) requires costly linear programming. The Sinkhorn algorithm (Cuturi, 2013) can estimate the WS distance fast by adding an entropy regularization, formulated as:

$$W(\alpha,\beta) = \min_{\boldsymbol{P}\in\Gamma(\alpha,\beta)}\left\{\langle \boldsymbol{P},\boldsymbol{D}\rangle - \eta\mathbb{E}(\boldsymbol{P})\right\}, \tag{7}$$

where $\mathbb{E}(\boldsymbol{P})$ is the entropy regularization, defined as:

$$\mathbb{E}(\boldsymbol{P}) = -\sum_{i=1,\ j=1}^{m,m}\boldsymbol{P}_{i,j}(\log\boldsymbol{P}_{i,j}-1), \tag{8}$$

and $\eta > 0$ is the entropy regularization weight. The optimization of Sinkhorn algorithm is realized by alternatively updating $\alpha = \boldsymbol{a}/\boldsymbol{S}\beta$ and $\beta = \boldsymbol{b}/\boldsymbol{S}\alpha$, where $\boldsymbol{S} = \exp(D)\in\mathbb{R}_+^{m\times m}$ and $/$ is the elementwise division. After updating the algorithm, one can finally obtain the optimal transportation plan $\boldsymbol{P}^* = \mathrm{diag}(\alpha^*)\boldsymbol{S}\mathrm{diag}(\beta^*)$.

Of note, the Sinkhorn distance has a unique optimal solution due to the strong concavity of the entropy regularization.

3.3.5. Symmetric optimal transport

Since the Sinkhorn distance can figure out an optimal WS distance for two distributions, it holds that the optimal transportation plan for $\mathrm{OT}(\alpha,\beta)$ and $\mathrm{OT}(\beta,\alpha)$ should be symmetric. However, in practice, we observe that the transportation plan for two OTs is non-symmetric, since the optimization of a DNN cannot reach the global optimum. It causes unstable performance of the DNN, leading to a significant performance drop. This is justified by the empirical study in Section 4. This issue also indicates that it is not good for OT to be a candidate as a metric in our specific application, in the scene where the symmetric property is essential for a valuable metric. Building on this, we believe the video-based VI-reID is a difficult task and it is a non-trivial contribution to bring performance gain via modifying the OT. We address this by developing symmetric optimal transport (SOT), formulated as:

$$\mathrm{SOT}(\alpha,\beta) := \frac{1}{2}\big(W(\alpha,\beta)+W(\beta,\alpha)\big). \tag{9}$$

The physical interpretation of the proposed SOT is to identify the matching flow in two directions, and it is obvious that the proposed SOT is a valid metric as it holds that $\mathrm{SOT}(\alpha,\beta) = \mathrm{SOT}(\beta,\alpha)$. We also empirically justify that the proposed SOT can work effectively as a way

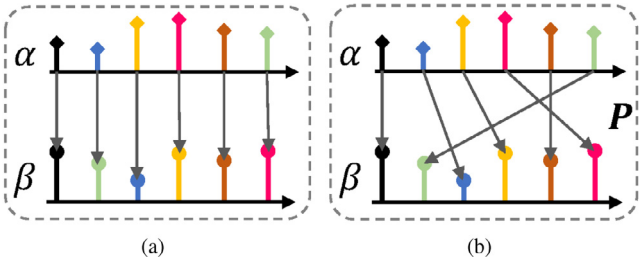


Fig. 4. Comparison of the distribution matching methods. (a) shows the matching paradigm of the most distribution measurement. The OT, shown in (b) realizes the optimal matching between two distributions.

to align two video sequences. In other words, the proposed SOT not only realizes a valid metric for our task but also offers a good practice of OT scheme in the challenging video-based VI-reID task. One can yield the approximation of SOT using Sinkhorn since it calculates the OT in two directions. Specifically, this is formulated as

$$\begin{aligned}\mathrm{SOT}(\alpha,\beta) &= \frac{1}{2}\big(W(\alpha,\beta)+W(\beta,\alpha)\big)\\ &= \min_{\boldsymbol{P}\in\Gamma(\alpha,\ \beta)}\left\{\langle \boldsymbol{P},\boldsymbol{D}\rangle - \eta\mathbb{E}(\boldsymbol{P})\right\}\\ &+ \min_{\boldsymbol{P}'\in\Gamma(\beta,\ \alpha)}\left\{\langle \boldsymbol{P}',\boldsymbol{D}'\rangle - \eta\mathbb{E}(\boldsymbol{P}')\right\}.\end{aligned} \tag{10}$$

Then each of the minimization process can be optimized individually via alternatively updating $(\alpha,\ \beta)$.

3.4. Optimization

In this paper, the discussed alignment methods are adopted as a component of the loss function, undertaking the role to optimize the neural network. We use the KLD as an example. Suppose $\alpha := \mathcal{V}_f = f\big(f(\mathcal{V}|\theta_v)|\theta_s\big)$ and $\beta := \mathcal{I}_f = f\big(f(\mathcal{I}|\theta_i)|\theta_s\big)$, the loss is given by $\mathcal{L} = \mathrm{KLD}(\alpha,\beta)$. The optimization is formulated as:

$$\theta_v^*,\theta_i^*,\theta_s^* = \underset{\theta_v,\ \theta_i,\ \theta_s}{\mathrm{argmin}}\big(\mathrm{KLD}(\alpha,\beta)\big). \tag{11}$$

Learning from the OT and SOT as a loss function can be formulated as a bi-level optimization problem. Given the loss $\mathcal{L} = \mathrm{OT}(\alpha,\beta)$ as an example, the bi-level optimization process can be formulated as follows:

$$\theta_v^*,\theta_i^*,\theta_s^* = \underset{\theta_v,\ \theta_i,\ \theta_s}{\mathrm{argmin}}\Big(\underset{\boldsymbol{P}}{\mathrm{argmin}}\big(\langle \boldsymbol{P},\boldsymbol{D}\rangle\big)\Big), \tag{12}$$

where $\boldsymbol{P}$ and $\boldsymbol{D}$ are transportation plan and distance matrix respectively.

Remark 1. This paper addresses the video VI-reID task by modeling the video sequence as a discrete distribution and aligning the distributions. In this section, we discuss the possible alignment distribution methods, that can be used as a loss constraint to optimize the network. The KLD, JSD, and MMD are very popular methods that aim to push two distributions close, without considering the matching flow of two distributions. In contrast to these "hard alignment" solutions (see Fig. 4(a)), OT becomes a reliable method that not only ensures numerical stability as a metric but also realizes an optimal matching flow between distributions. This suggests employing this "soft alignment" solution as an alternative method to matching the video sequences (see Fig. 4(b)).

4. Experiments

4.1. Datasets and evaluation metric

Video VI-reID is a new task and only the HITSZ Video Cross-Modal (HITSZ-VCM) dataset is available as the benchmark for this

(a) Samples of the video sequence in visible modality. (b) Samples of the video sequence in infrared modality.

Fig. 5. Video frames sampled from the pedestrian video sequences in the HITSZ-VCM dataset. These six video sequences, captured by different cameras, share the same person identity.

task, thus all studies throughout this paper are conducted on the HITSZ-VCM dataset. This dataset contains 251,452 visible images and 211,807 infrared images, constructing 11,785 and 10,078 video tracklets. Each modality of video data has 927 person identities. Multiple non-overlapped cameras are employed to capture each person in a different view, providing a rich appearance for each person. This dataset is captured by 12 HD cameras and covers a series of scenes, e.g., 7 outdoor scenes, 3 indoor scenes, and 2 passages scenes. Those scenes include the office, cafe, garden, etc. Fig. 5 illustrates the samples of video sequences captured by different cameras. In the inference phase, a trained neural network is evaluated by two retrieval modes, e.g., Infrared to Visible mode and Visible to Infrared mode.

In person re-ID task, two popular metrics are commonly used to evaluate the performance of algorithms, e.g., cumulative matching characteristic (CMC) curve and mean average precision (mAP). For a given query sample, the CMC curve shows the correct matching rate at various ranks, whereas the mAP value indicates the overall ranking performance. In this paper, both two metrics are adopted to evaluate the video-based VI-re-ID machine.

4.2. Implementation details

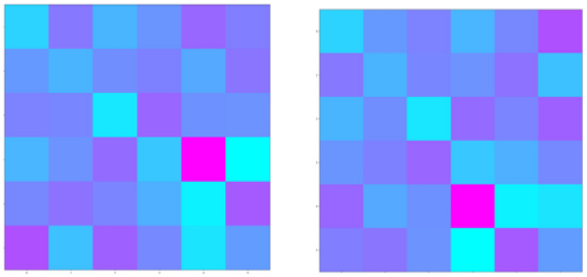
We implement our algorithm in the PyTorch (Paszke et al., 2017) and all experiments are performed on NVIDIA 3090 GPUs. We use MITM (Lin et al., 2022) as a baseline to evaluate our method. For each video clip, we $T = 6$ in all experiments. Each frame is resized to 288×144. The data augmentations used in our experiments include zero-padding and randomly flipping in the horizontal direction. The batch size is set to 8 for each modality of video data. The network is optimized by SGD optimizer, where the weight decay and momentum are set to 5×10^{-4} and 0.9. The warmup strategy is also employed for the learning rate adaptation. The learning rate is initialized to 0.1, and decayed at the 35th and 80th epochs with a factor of 0.1.

4.3. Ablation study

In this part, we first conduct extensive studies to explore the benefits of the proposed symmetric optimal transport in the video VI-reID task.

4.3.1. Analysis of the non-symmetric OT

Given two video sequences with each belonging to one modality, one can identify an optimal transportation plan, realizing a minimal matching cost in the raw image space. To align the cross-modality data, a non-linear function is employed to project the video data to an embedding space. That said, the same transportation plan can be identified in the embedding space, resulting in a minimal matching cost. It holds that $\boldsymbol{P}^*_{\text{img}} = \boldsymbol{P}^*_{\text{feat}}$, where $\boldsymbol{P}^*_{\text{img}} = \text{argmin}_{\boldsymbol{P}_{\text{img}}}\big(\text{OT}(\mathcal{V}, \mathcal{I})\big)$ and $\boldsymbol{P}^*_{\text{feat}} = \text{argmin}_{\boldsymbol{P}_{\text{feat}}}\big(\text{OT}(\mathcal{V}_f, \mathcal{I}_f)\big)$. In this optimal solution, the Wasserstein distance should be symmetric in the embedding space, e.g., $W(\mathcal{V}_f, \mathcal{I}_f) = W(\mathcal{I}_f, \mathcal{V}_f)$.



(a) Optimal transportation plan for $W(\mathcal{V}_f, \mathcal{I}_f)$. (b) Optimal transportation plan for $W(\mathcal{I}_f, \mathcal{V}_f)$.

Fig. 6. The visualization of the transportation plan for the $W(\mathcal{V}_f, \mathcal{I}_f)$ and $W(\mathcal{I}_f, \mathcal{V}_f)$ using a common feature extractor. In (a) and (b), the color per element indicates the value of the transportation plan, and a darker color indicates a larger value. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Evaluation of the modality learning network on HITSZ-VCM dataset. We use the bold to indicate the best result.

Models	Infrared to Visible		Visible to Infrared	
	R-1	mAP	R-1	mAP
w/ modality learning	**63.74**	45.31	**64.54**	**47.69**
w/o modality learning	58.27	40.38	60.02	43.24

Table 2
Comparison of the performance of $\text{OT}(\alpha, \beta)$ and $\text{OT}(\beta, \alpha)$ on HITSZ-VCM dataset.

Models	Infrared to Visible		Visible to Infrared	
	R-1	mAP	R-1	mAP
Baseline	**63.74**	45.31	64.54	47.69
$\text{OT}(\alpha, \beta)$	62.61	45.24	**65.46**	**47.88**
$\text{OT}(\beta, \alpha)$	63.02	**46.06**	61.60	41.89

Theoretically, once two modalities of video are encoded by the same embedding function (e.g., $f(\cdot|\theta_v) = f(\cdot|\theta_i)$), the Wasserstein distance is symmetric, leading to a symmetric transportation plan, as shown in Fig. 6. However, the video VI-reID task is a challenging task and recent studies show that the modality learning networks, e.g., $f(\cdot|\theta_v)$ and $f(\cdot|\theta_i)$ in Fig. 3, are required to learn the low-level domain knowledge per modality. As shown in Table 1, the modality learning network contributes considerably to the feature embedding function in the video VI-reID task (Ye et al., 2021b, 2020).

Even though this modality learning network plays a vital role in the feature embedding, it raises an issue that such a neural architecture (see Fig. 3) results in a non-symmetric transportation plan, as observed in Fig. 2. This issue results in the unstable performance of the network. Specifically, for two distributions α and β, we optimize the network by either $\text{OT}(\alpha, \beta)$ or $\text{OT}(\beta, \alpha)$, and the result is reported in Table 2. It shows that both $\text{OT}(\alpha, \beta)$ and $\text{OT}(\beta, \alpha)$ have difficulties bringing performance gain consistently, and adopting the OT directly cannot improve the performance, again showing that the video-based VI-reid is a difficult problem.

4.3.2. Evaluation on different alignment methods

In the above part, we analyze that vanilla OT cannot be adopted in the existing SOTA models. This part continues to show the proposed SOT can attain superior performance than the well-established distribution alignment methods. Specifically, we employ the Kullback–Leibler divergence (KLD), Jensen–Shannon divergence (JSD), maximum mean discrepancy (MMD), vanilla optimal transport (OT), and the proposed

Table 3
Comparison with different distribution alignment methods on HITSZ-VCM dataset. We use the bold to indicate the best result.

Models	Infrared to Visible		Visible to Infrared	
	R-1	mAP	R-1	mAP
Baseline	63.74	45.31	64.54	47.69
KLD	63.98	44.81	64.50	47.24
JSD	61.65	45.00	64.82	48.22
MMD	64.26	45.92	65.22	47.80
OT	62.61	45.24	65.46	47.88
SOT	**64.97**	**47.74**	**67.93**	**49.67**

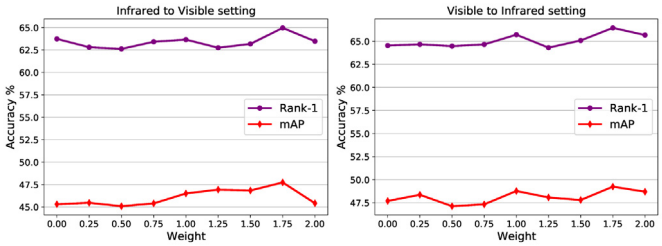


Fig. 7. Comparison with different weights of symmetric optimal transport loss on HITSZ-VCM dataset. This study is evaluated on both "Visible to Infrared" and "Infrared to Visible" retrieval settings.

symmetric optimal transport (SOT), in this study. It is noted that those methods are used as a component in the loss function. The result is reported in Table 3. We can find that those existing methods cannot bring consistent performance gain over the two retrieval modes, e.g., infrared to visible and visible to infrared modes. This shows that video VI-reID is a challenging task. Also noted, the vanilla OT only brings performance gain on the "Visible to Infrared" mode, showing the unstable inference from the network. In the meantime, this study also reveals that the proposed SOT can improve the baseline network consistently, clearly showing the superiority of the proposed method. As compared to the vanilla OT method, our method demonstrates its effectiveness, justifying that we make a good practice of distribution alignment using the SOT scheme.

It is noted that the simple MMD can achieve an overall best performance than the well-establish "measurement". However, the proposed SOT is better than the MMD, especially in the "Visible to Infrared" mode, where the SOT outperforms the MMD by 2.71%/1.87% on R-1/mAP. This shows the superiority of the proposed SOT.

We also provide the difference of the intra/inter-class variance to interpret the performance gain of our method. Specifically, given the similarity distribution of positive and negative pairs, the intra/inter-class variance is the distance of medium value between two distributions. Our method can improve the distance by 0.24 over that of the baseline, showing that our method can enlarge the margin between intra/inter-class variance.

4.3.3. Evaluation on different weights

In our practice, the distribution alignment method is employed as a component of the loss function, thereby constraining the parameter updating in the back-propagation procedure. In this context, one needs to identify a proper value to weigh the loss component. In doing so, we conduct an empirical study to choose the weight value as shown in Fig. 7. Figs. 7(a) and 7(b) illustrate the retrieval performance w.r.t. to different weight values in "Infrared to Visible" setting and "Visible and Infrared" setting, respectively. It shows that in both settings, the network achieves the best performance when the weight value is 1.75. Throughout this paper, the weight value for the SOT is set to 1.75.

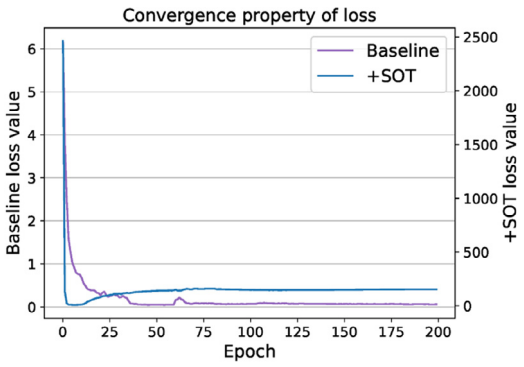


Fig. 8. The convergence property of the loss function. We compare the convergence curves of loss from baseline and our method.

4.3.4. Convergence property of loss

In this part, we study the convergence property of the loss. Specifically, we visualize the convergence curves of the loss functions for baseline vs. +SOT, as shown in Fig. 8. It shows that even though the loss value of SOT is very large, it convergences faster than the loss in the baseline network, indicating that it does not increase the training difficulties to optimize the network using the proposed SOT and the proposed scheme for distribution measurement is suit for our task.

4.3.5. Evaluation on various baseline networks

We continue to study the generalization of the proposed SOT by evaluating its effectiveness on various baseline networks. Specifically, three baselines, e.g., VI-ResNet-50, VI-GLTR, MITM, are adopted in this study. All baselines are pre-trained on ImageNet (Russakovsky et al., 2015). The VI-ResNet-50 is derived from the ResNet-50 (He et al., 2016). In doing so, the convolutional block, i.e., conv1, is instantiated to low-level modality learning networks, i.e., $f(\cdot|\theta_v)$ and $f(\cdot|\theta_i)$. The following residual blacks, e.g., conv2_x to conv5_x, are instantiated to the high-level feature learning network, e.g., $f(\cdot|\theta_s)$. Then the proposed SOT is further adopted as one of the loss components to optimize the network. The GLTR, which is short for global-local temporal representation, is first proposed for the video person reID task (Li et al., 2019). To adapt the GLTR for the video VI-reID task, a minor modification is required. To this end, the backbone network of the GLTR, e.g., ResNet-50, should be split into two parts, a low-level modality learning network, and a high-level feature learning network, as the modification in VI-ResNet-50 discussed above. The most SOTA method, MITM, is proposed for the video VI-reID task, such that it can be used to evaluate the proposed SOT directly.

Fig. 9 illustrates the improvement of the SOT on top of various baselines. It shows that the proposed SOT scheme can bring consistent improvements in "Infrared to Visible" retrieval mode (see Figs. 9(a) and 9(b)) and "Visible to Infrared" retrieval mode (see Figs. 9(c) and 9(d)) across all baselines. It reveals that the proposed SOT is an effective yet flexible method.

4.3.6. Evaluation on the robustness

We further show the robustness property of our method. Table 4 indicates that the performance of the network incorporated with the optimal transport (OT) is unstable since the performance of OT(α, β) and OT(β, α) various significantly, while the proposed SOT attains stable performance for SOT(α, β) and SOT(β, α), showing the superiority and robustness of the proposal.

To further evaluate its robustness, we evaluate the methods using noise labels. Specifically, we add 20% noise to the label. The performance drop w.r.t. R-1/mAP value of the baseline and our method reads as 10.62%/8.62% vs. 6.24%/5.80%, clearly showing the robustness of our method.

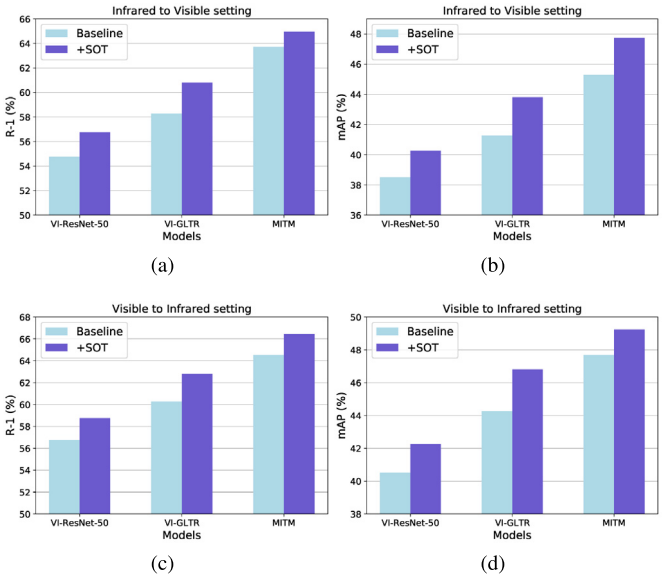


Fig. 9. Evaluation of the proposed symmetric optimal transport on various baseline networks. This study is evaluated on both "Visible to Infrared" and "Infrared to Visible" retrieval settings.

Table 4
Comparison of the robustness of OT and SOT on HITSZ-VCM dataset.

Models	Infrared to Visible		Visible to Infrared	
	R-1	mAP	R-1	mAP
Baseline	63.74	45.31	64.54	47.69
OT(α,β)	62.61	45.24	65.46	47.88
OT(β,α)	63.02	46.06	61.60	41.89
SOT(α,β)	64.97	47.74	67.93	49.67
SOT(β,α)	64.82	47.48	68.12	50.21

4.4. Comparison to state-of-the-art methods

In Section 4.3, extensive studies are conducted to evaluate the effectiveness of the proposed method. To further evaluate its superiority, we compare our method to the SOTA models. In doing so, we compare our method to LbA (Park et al., 2021), MPANet (Wu et al., 2021), DDAG (Ye et al., 2020), VSD (Tian et al., 2021), CAJL (Ye et al., 2021a) and MITM (Lin et al., 2022). LbA addresses the alignment issue by exploiting the dense correspondence between cross-modal person images, such that suppressing the domain features of two modalities (Park et al., 2021). In MPANet, the alignment is realized by discovering cross-modality nuances, following two steps: dislodging the modality information and aligning the patterns (Wu et al., 2021). In DDAG, both the part level and contextual level are leveraged to jointly learn discriminative person representation and mitigate the modality effects (Ye et al., 2020). Learning the representation from the information bottleneck view, VSD creates the embedding space by fitting the mutual information (Tian et al., 2021). Considering the imagery property of the visible image and infrared image, CAJL develops a data augmentation to homogeneously generate color-irrelevant images by randomly exchanging the color channels. It is noted that these methods are initially developed for the image-level VI-reID task. The most recent SOTA model, i.e., MTIM, is tailored for the video VI-reID task and the property of the video data motivates to develop of a temporal-memory mechanism, that produces the motion-invariant embedding for the video data (Wu et al., 2021).

The comparison is made in Table 5. It shows that our work outperforms existing SOTA methods and attains a new SOTA performance in the video VI-reID task. For example, in the "Infrared to Visible" retrieval mode, our method outperforms the MITM by 1.23%/2.43% on R-1/mAP values. In another mode, our method also outperforms the MITM and the improvement reads as 3.39%/1.98% on R-1/mAP values. This indeed shows the superiority of our proposal in this paper.

4.5. Evaluation on other applications

The SOT is proposed to identify the optimal matching flow between two distributions in two directions. That said, one can model any modality of data to a distribution and adopt the proposed method to align any two distributions. Along with the evaluation of video-based VI-reID in the main paper, we further evaluate the effectiveness of the proposed SOT on the image-based VI-reID per the comment. Specifically, our evaluation is conducted on standard benchmarks, *i.e.*, SYSU-MM01 (Wu et al., 2017) and RegDB (Nguyen et al., 2017) datasets. We follow the common practice to evaluate our method in both R-1 and mAP values (Ye et al., 2020). The result is reported in Tables 6 and 7. It shows that the proposed SOT can work effectively in the image-based VI-reID task and bring consistent performance gain over the baseline on both SYSU-MM01 and RegDB datasets.

We also evaluate the effectiveness of the proposed SOT by comparing it with classic VI-reID methods including FBP-AL (Wei et al., 2021a), NFS (Chen et al., 2021), RBDF (Wei et al., 2022), CM-NAS (Fu et al., 2021), SMCL (Wei et al., 2021b) and CAJL (Ye et al., 2021a). The comparison is made in Tables 6 and 7. It shows that our method attains a very competitive performance compared to the SOTA methods. Especially in the RegDB dataset, our method can significantly improve the SOTA performance, vividly showing the superior property and generalization of our method.

5. Conclusion

In this paper, we study the video-based visible-infrared person re-identification task. This is a video-matching problem with heterogeneous data formats. Either learning modality-invariant person features or exploring the temporal information in the video embedding can be an option to address this task. This paper considers addressing the video-matching problem via determining the optimal matching flow between videos. In doing so, we formulate the video clip as a discrete distribution formally and align the distributions with divergence or distance metrics. Empirical studies show that the existing well-established distribution alignment methods, including the optimal transport, have

Table 5
Comparison with the state-of-the-art algorithms on HITSZ-VCM dataset. We use the bold to indicate the best result.

Models	Venue	Infrared to Visible					Visible to Infrared				
		R-1	R-5	R-10	R-20	mAP	R-1	R-5	R-10	R-20	mAP
LbA (Park et al., 2021)	ICCV'21	46.38	65.29	72.23	79.41	30.69	49.30	69.27	75.90	82.21	32.38
MPANet (Wu et al., 2021)	CVPR'21	46.51	63.07	70.51	77.77	35.26	50.32	67.31	73.56	79.66	37.80
DDAG (Ye et al., 2020)	ECCV'20	54.62	69.79	76.05	81.50	39.26	59.03	74.64	79.53	84.04	41.50
VSD (Tian et al., 2021)	CVPR'21	54.53	70.01	76.28	82.01	41.18	57.52	73.66	79.38	83.61	43.45
CAJL (Ye et al., 2021a)	ICCV'21	56.59	73.49	79.52	84.05	41.49	60.13	78.96	82.98	87.10	42.81
MITM (Lin et al., 2022)	CVPR'22	63.74	76.88	81.72	86.28	45.31	64.54	78.96	82.98	87.10	47.69
Ours	This work	**64.97**	**78.12**	**82.81**	**86.87**	**47.74**	**67.93**	**81.07**	**84.94**	**88.59**	**49.67**

Table 6
Comparison with the state-of-the-art algorithms on SYSU-MM01 dataset.

Models	All search		Indoor search	
	R-1	mAP	R-1	mAP
FBP-AL (Wei et al., 2021a)	54.14	50.20	73.98	50.20
NFS (Chen et al., 2021)	56.91	55.45	62.79	69.79
RBDF (Wei et al., 2022)	57.66	54.41	–	–
CM-NAS (Fu et al., 2021)	61.99	60.02	67.01	72.95
SMCL (Wei et al., 2021b)	67.39	61.78	68.84	75.56
CAJL (Ye et al., 2021a)	**69.88**	**66.89**	76.26	**80.37**
Baseline	66.71	64.71	72.73	77.63
+ SOT	68.72	66.80	**76.96**	79.64

Table 7
Comparison with the state-of-the-art algorithms on RegDB dataset.

Models	Visible to Infrared		Infrared to Visible	
	R-1	mAP	R-1	mAP
NFS (Chen et al., 2021)	80.54	72.10	77.95	69.79
FBP-AL (Wei et al., 2021a)	73.98	68.24	70.05	66.61
RBDF (Wei et al., 2022)	79.80	76.71	76.21	73.92
SMCL (Wei et al., 2021b)	83.93	79.83	83.05	78.57
CM-NAS (Fu et al., 2021)	84.54	80.32	82.57	78.31
CAJL (Ye et al., 2021a)	85.03	79.14	84.75	77.82
Baseline	90.57	83.05	90.32	82.95
+ SOT	**92.76**	**86.07**	**93.67**	**86.30**

difficulties bringing performance gain consistently over different retrieval modes. Given this, we further propose a symmetric optimal transport scheme as a better way to align distributions. Thorough empirical results justify the effectiveness of the proposed methods. We believe our idea will be a promising direction for the video-matching problem. Even though the proposed method reports the SOTA result, the retrieval accuracy is still low, and more effort is required to develop new methods.

In the future, we will work on alternative solutions, which consider the geometric distribution of video frames, to align the video data, and improve the retrieval performance of the network.

CRediT authorship contribution statement

Pengfei Fang: Conceptualization, Methodology, Software, Writing. **Yaojun Hu:** Software, Investigation. **Shipeng Zhu:** Software, Investigation, Reviewing. **Hui Xue:** Supervision, Reviewing and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62306070, 62076062 and in part by the Southeast University Start-Up Grant for New Faculty under Grant 4009002309. Furthermore, the work was also supported by the Big Data Computing Center of Southeast University.

References

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein GAN. arXiv:1701.07875 [stat.ML].

Bellemare, M.G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., Munos, R., 2017. The cramer distance as a solution to biased wasserstein gradients. arXiv:1705.10743 [cs.LG].

Chen, Y., Wan, L., Li, Z., an Zongyuan Sun, Q.J., 2021. Neural feature search for RGB-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 587–597.

Cho, Y.-J., Kim, S.-A., Park, J.-H., Lee, K., Yoon, K.-J., 2019. Joint person re-identification and camera network topology inference in multiple cameras. Comput. Vis. Image Underst. 180, 34–46.

Choi, S., Lee, S., Kim, Y., Kim, T., Kim, C., 2020. Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10257–10266.

Cuturi, M., 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In: Proceedings of the Advances in Neural Information Processing Systems, Vol. 26. pp. 2292–2300.

del Barrio, E., Inouzhe, H., Loubes, J.-M., Matran, C., Mayo-Iscar, A., 2020. OptimalFlow: optimal transport approach to flow cytometry gating and population matching. BMC Bioinformatics 21, 1–25.

Fang, P., Zhou, J., Roy, S.K., Ji, P., Petersson, L., Harandi, M., 2021. Attention in attention networks for person retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 44, 4626–4641.

Fang, P., Zhou, J., Roy, S.K., Petersson, L., Harandi, M., 2019. Bilinear attention networks for person retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8030—8039.

Fu, C., Hu, Y., Wu, X., Shi, H., Mei, T., He, R., 2021. CM-NAS: Cross-modality neural architecture search for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11823–11832.

Genevay, A., Peyré, G., Cuturi, M., 2017. Learning generative models with sinkhorn divergences. arXiv:1706.00292 [stat.ML].

Gong, S., Cristani, M., Yan, S., Loy, C.C., 2014. Person Re-Identification. Springer.

Gong, J., Zhao, S., Lam, K.-M., Gao, X., Shen, J., 2023. Spectrum-irrelevant fine-grained representation for visible–infrared person re-identification. Comput. Vis. Image Underst. 232, 103703.

Hao, X., Zhao, S., Ye, M., Shen, J., 2021. Cross-modality person re-identification via modality confusion and center aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16403–16412.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hermans, A., Beyer, L., Leibe, B., 2017. In defense of the triplet loss for person re-identification. arXiv:1703.07737 [cs.CV].

Kolouri, S., Naderializadeh, N., Gustavo K, R., Hoffmann, H., 2021. Wasserstein embedding for graph learning. In: Proceedings of the International Conference on Learning Representations. pp. 1–20.

Li, Z., Lv, J., Chen, Y., Yuan, J., 2021. Person re-identification with part prediction alignment. Comput. Vis. Image Underst. 205, 103172.

Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S., 2019. Global-local temporal representation for video person re-identification. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. pp. 3958–3967.

Li, D., Wei, X., Hong, X., Gong, Y., 2020. Infrared-visible cross-modal person re-identification with an X modality. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 4610–4617.

Lin, X., Li, J., Ma, Z., Li, H., Li, S., Xu, K., Lu, G., Zhang, D., 2022. Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20973–20982.

Liu, J., Liu, J., Zhang, Q., 2023a. M2FINet: Modality-specific and modality-shared features interaction network for RGB-IR person re-identification. Comput. Vis. Image Underst. 232, 103708.

Liu, Z., Mu, X., Lu, Y., Zhang, T., Tian, Y., 2023b. Learning transformer-based attention region with multiple scales for occluded person re-identification. Comput. Vis. Image Underst. 229, 103652.

Liu, Y., Zhu, L., Yamada, M., Yang, Y., 2020. Semantic correspondence as an optimal transport problem. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4463–4472.

Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N., 2020. Cross-modality person re-identification with shared-specific feature transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13379–13389.

Monge, G., 1781. Mémoire Sur La Théorie Des Déblais Et Des Remblais. Histoire de l'Académie Royale des Sciences de Paris, pp. 666–704.

Nguyen, D.T., Hong, H.G., Kim, K.W., Park, K.R., 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors 17, 1–29.

Park, H., Lee, S., Lee, J., Ham, B., 2021. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12046–12055.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in PyTorch. In: Proceedings of the Thirty-First Conference on Neural Information Processing Systems. pp. 1–4.

Rubner, Y., Tomasi, C., Guibas, L.J., 2000. The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. 40, 99–121.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., Li, F., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.

Salimans, T., Zhang, H., Radford, A., Metaxas, D., 2018. Improving GANs using optimal transport. arXiv:1803.05573 [cs.LG].

Schroff, F., Kalenichenko, D., Philbin, J., 2015. FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. pp. 815–823.

Solomon, J., Rustamov, R., Guibas, L., Butscher, A., 2014. Earth mover's distances on discrete surfaces. ACM Trans. Graph. 33, 1–12.

Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q., 2017. Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3960–3969.

Suh, Y., Wang, J., Tang, S., Mei, T., Mu Lee, K., 2018. Part-aligned bilinear representations for person re-identification. In: Proceedings of the European Conference on Computer Vision. pp. 418–437.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S., 2018. Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision. pp. 501–518.

Tay, C.-P., Roy, S., Yap, K.-H., 2019. AANet: Attribute attention network for person re-identifications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7134–7143.

Tian, X., Zhang, Z., Lin, S., Qu, Y., Ma, Y.X.L., 2021. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1522–1531.

Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.-Y., Satoh, S., 2019b. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 618–626.

Wang, Z., Wang, Z., Zheng, Y., Wu, Y., Zeng, W., Satoh, S., 2020b. Beyond intra-modality: A survey of heterogeneous person re-identification. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Survey Track. pp. 4973–4980.

Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z., 2019a. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3623–3632.

Wang, G.-A., Zhang, T., Yang, Y., Cheng, J., Chang, J., Liang, X., Hou, Z., 2020a. Cross-modality paired-images generation for RGB-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 12144–12151.

Wang, J., Zhou, F., Wen, S., Liu, X., Lin, Y., 2017. Deep metric learning with angular loss. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2593–2601.

Wei, X., Li, D., Hong, X., Ke, W., Gong, Y., 2020. Co-attentive lifting for infrared-visible person re-identification. In: Proceedings of the ACM International Conference on Multimedia. pp. 1028–1037.

Wei, Z., Yang, X., Wang, N., Gao, X., 2021a. Flexible body partition-based adversarial learning for visible infrared person re-identification. IEEE Trans. Neural Netw. Learn. Syst. 33, 4676–4687.

Wei, Z., Yang, X., Wang, N., Gao, X., 2021b. Syncretic modality collaborative learning for visible infrared person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 225–234.

Wei, Z., Yang, X., Wang, N., Gao, X., 2022. RBDF: Reciprocal bidirectional framework for visible infrared person reidentification. IEEE Trans. Cybern. 52, 10988–10998.

Weinberger, K.Q., Saul, L.K., 2009. Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. 10, 207–244.

Wu, Q., Dai, P., Chen, J., Lin, C.-W., Wu, Y., Huang, F., Zhong, B., Ji, R., 2021. Discover cross-modality nuances for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4330–4339.

Wu, A., Zheng, W.-S., Gong, S., Lai, J., 2020. RGB-IR person re-identification by cross-modality similarity preservation. Int. J. Comput. Vis. 1765—1785.

Wu, A., Zheng, W.-S., Yu, H.-X., Gong, S., Lai, J., 2017. Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5380–5389.

Xu, H., Luo, D., Zha, H., Duke, L.C., 2019. Gromov-wasserstein learning for graph matching and node embedding. In: Proceedings of the 36th International Conference on Machine Learning, Vol. 97. pp. 6932–6941.

Ye, M., Ruan, W., Du, B., Shou, M.Z., 2021a. Channel augmented joint learning for visible-infrared recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13567–13576.

Ye, M., Shen, J., Crandall, D.J., Shao, L., Luo, J., 2020. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: Proceedings of the European Conference on Computer Vision. pp. 229–247.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H., 2021b. Deep learning for person re-identification: A survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. 44, 2872–2893.

Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F., Solomon, J.M., 2019. Hierarchical optimal transport for document representation. In: Proceedings of the Advances in Neural Information Processing Systems, Vol. 32. pp. 1601–1611.

Zhang, C., Cai, Y., Lin, G., Shen, C., 2020. DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12203–12213.

Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H., 2018. Deep mutual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4320–4328.

Zhao, H., Phung, D., Huynh, V., Le, T., Buntine, W., 2021. Neural topic model via optimal transport. In: Proceedings of the International Conference on Learning Representations. pp. 1–11.

Zheng, L., Yang, Y., Hauptmann, A.G., 2016. Person re-identification: Past, present and future. arXiv:1610.02984 [cs.CV].

Zhou, X., Zhong, Y., Cheng, Z., Liang, F., Ma, L., 2017. Adaptive sparse pairwise loss for object re-identification. arXiv:2303.18247 [cs.CV].